Assessing the completeness of immunogenetic databases across European Populations

Shaunak Kapur, Yu Ning Huang, Serghei Mangul

Dept of Clinical Pharmacy and Quantitative and Computational Biology, Bridge Institute, University of Southern California, Los Angeles, CA , USA

Bridge UnderGrad Science (BUGS) ***Summer Research Program***

USC University of Southern California

Abstract

Studies that are traditionally focused on T-cell Receptors (TCR) within the immune systems rely on databases that are primarily comprised of Eurocentric data. However, the thoroughness and diversity within these datasets are not known. By analyzing the datasets we can gain valuable insights into phenotypic variations and diverse human responses to immune-related diseases. This inclusive approach will not only enrich our understanding of immunogenetics within European populations but also contribute to a more equitable representation of immune responses in this region.

Objectives

Developed the bioinformatics pipeline to examine the completeness of the (IMGT) database for European based ancestry groups and applied the pipeline to assess the two TCR-Seq studies

Methodologies

- Selected 2 human TCR-Seq studies with available TCR-Seq in the Sequence Read Archives (SRA): SRP073308 (11 samples), SRP028752 (4 studies)
- Downloaded the fastq files of the TCR-Seq from SRA via SRA-Toolkit
- Used MiXCR to align the sample TCR-Seq to the IMGT database and export the results
- Used specific Jupyter notebook in python to assess the completeness of the IMGT database in representing different studies based on the number of mismatches
- Counted and calculated the number of mismatches including substitution, insertion and deletion in the V gene across European ancestries to assess the representativeness of the studies to the IMGT database.

Figures 1 and 2

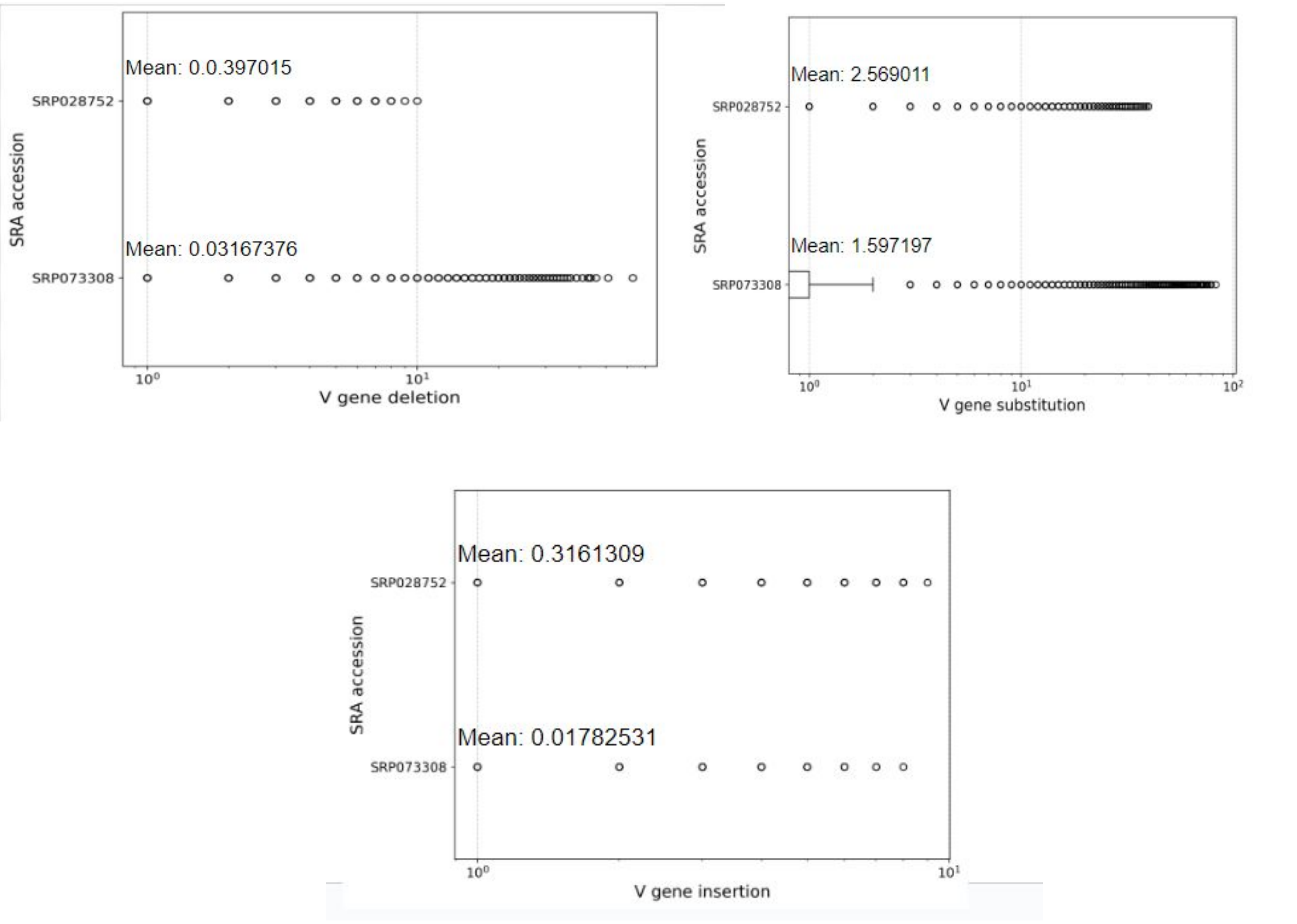

Figure 1. The number of the substitutions, insertions and deletions of the V genes among the European samples across two TCR-seq studies

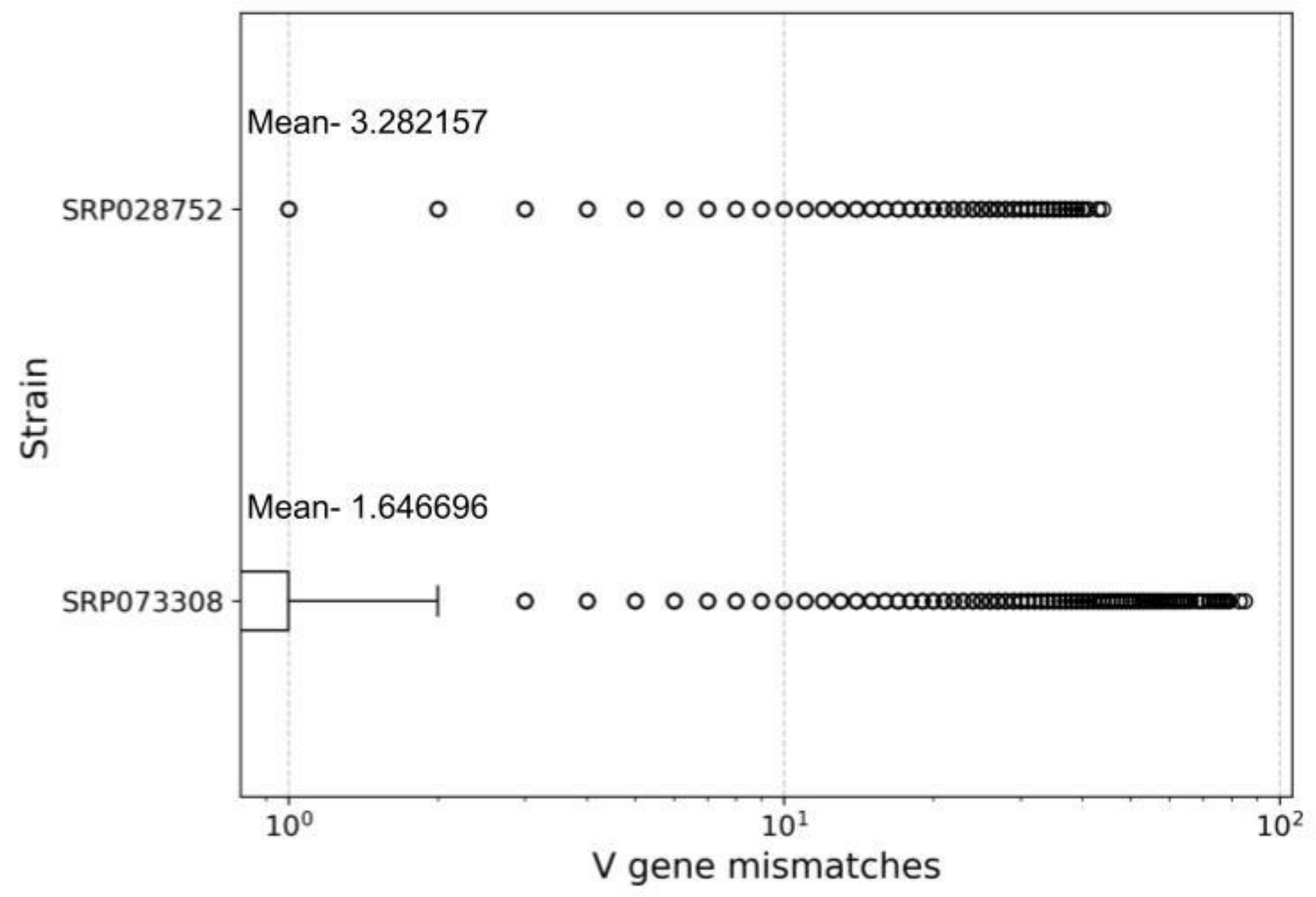

Figure 2. The total number of mismatches, including substitutions, insertions and deletions, of the V genes among the European Samples.

Figure 3

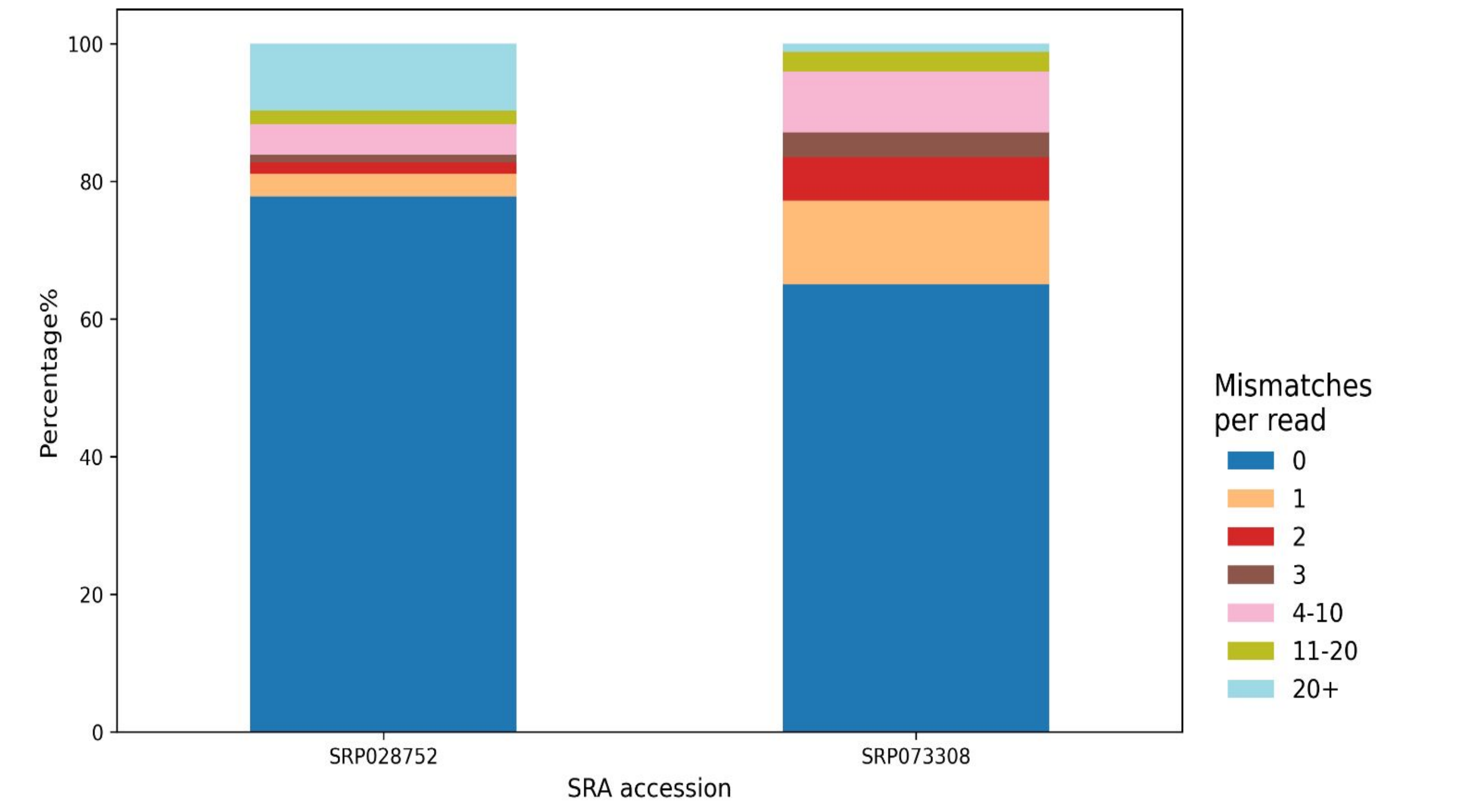

Figure 3. The percentage of the reads in the two studies with no mismatches, 1,2,3,4 to 10, 11 to 20 mismatches, and over 20 mismatches.

Conclusions

- Based on the data provided, a majority of the samples have 0 mismatches compared to the IMGT database, meaning the IMGT database is able to capture those alleles in such samples.
- The completeness of IMGT database in presenting the samples in SRP028752 studies is 77.8%, while in SRP073308 is 65.0%.
- The reads with over 20 mismatches are likely to be not captured by the IMGT database, and are likely to be the novel alleles that are not recorded in the IMGT database.
- SRP028752 is more representative of the V genes in the IMGT database that SRP073308
- Further examination is required to assess the completeness of the IMGT database in representing diverse ancestries.

Acknowledgements

Thank you to my mentor Mr. Huang for guiding me through this project process, and Dr. Mangul for enabling me to complete this opportunity through his lab!

CONTACT US

bridge.usc.edu/bugs